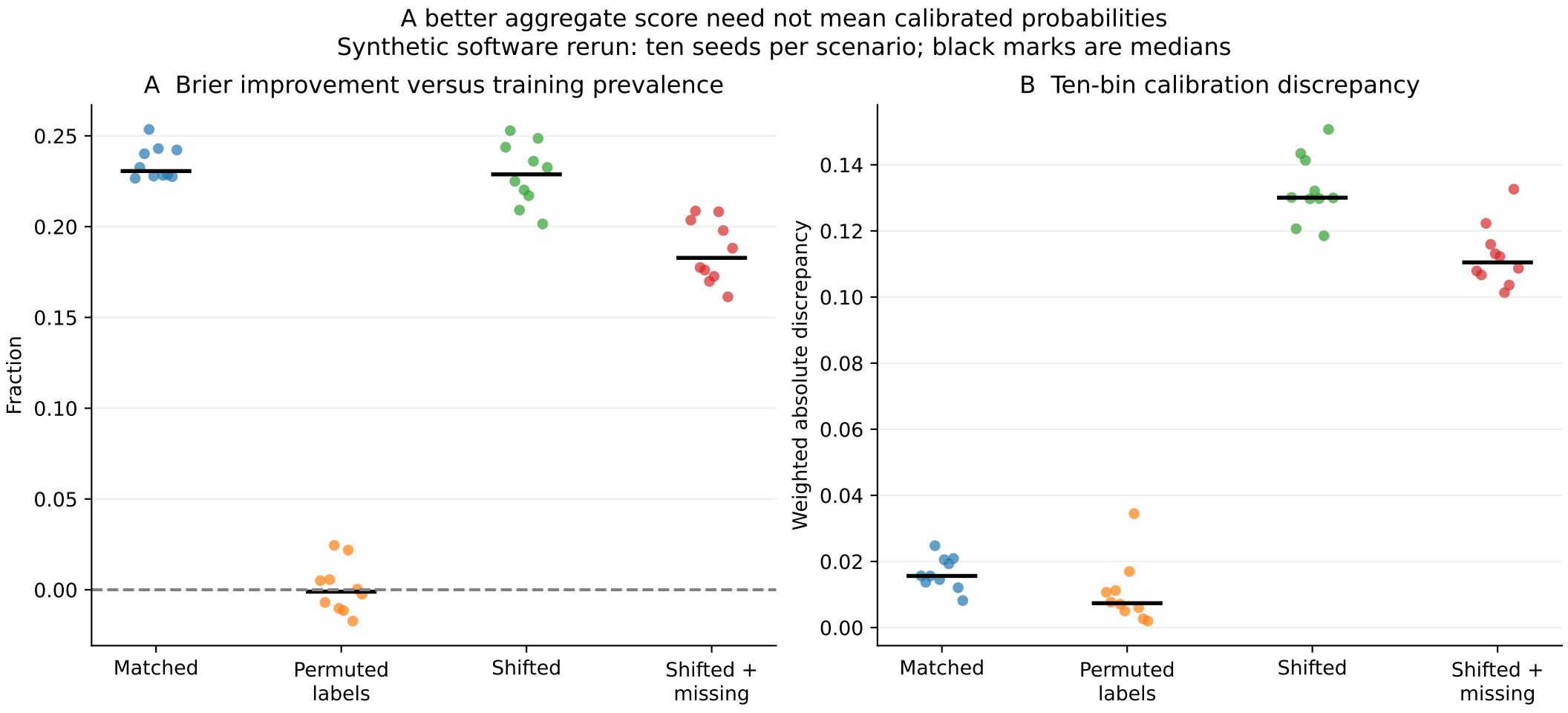

A better aggregate score need not mean calibrated probabilities
Synthetic software rerun: ten seeds per scenario; black marks are medians
A Brier improvement versus training prevalence
B Ten-bin calibration discrepancy
Fraction
Weighted absolute discrepancy
0.25
0.20
0.15
0.10
0.05
0.00
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Matched
Permuted labels
Shifted
Shifted + missing